Bayesian Music Alignment(Abstract_要旨)

AUTHOR(S):

Maezawa, Akira

CITATION:

Maezawa, Akira. Bayesian Music Alignment. 京都大学, 2015, 博士(情報学)

ISSUE DATE:

2015-03-23

URL:

https://doi.org/10.14989/doctor.k19106

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	前澤　陽
論文題目	Bayesian Music Alignment （ベイズ推定に基づく音楽アライメント）		

（論文内容の要旨）

This thesis addresses temporal alignment of music audio signals with a symbolic music score and alignment of multiple music audio signals, each of which represents a common piece of music (music alignment). This is an important task in many fields such as music production, musicological analysis, and informed sound source separation. This thesis focuses on alignment among multiple audio signals (audio-to-audio alignment), and that between a symbolic score and one or more audio signals (audio-to-score alignment). Moreover, the thesis focuses on Western music, which is polyphonic, consists mostly of harmonic sounds, and is played according to a music score.

Music alignment is difficult because there are variations among different performances, even though they all play a same music score. Such a variation of music audio signals calls for a probabilistic treatment of music audio signals. Furthermore, the fact that different audio signals have some aspects in common calls for a framework for encoding different constraints that are known in advance. These requirements can be satisfied through Bayesian inference. This thesis approaches music alignment from a Bayesian perspective. Chapter 1 presents an overview of music alignment and its relevance. Chapter 2 reviews background on alignment techniques.

Chapter 3 presents a Bayesian audio-to-score alignment method. Here, the symbolic music score is given, and the main goal is to infer the alignment, taking into account variations of audio signals in timbre, volume, and tempo. To deal with variations in timbre and dynamics, polyphonic pitched sounds in the spectral domain are modeled as a probabilistically weighted sum of spikes. These spikes are placed at frequencies where prominent energy is expected to be observed, via analyzing which notes are present in the music score at a given position. Robustness to variations in timbre and dynamics is attained by assuming a weakly informative prior on the weights. Furthermore, the model represents a smoothly changing tempo trajectory shared among different parts, while allowing for slight asynchronies between different parts. An experimental evaluation demonstrated that the proposed method achieved a median alignment error of about 30 ms on real-world ensemble pieces, and about 50 ms on synthetic orchestral recordings.

Chapter 4 presents a Bayesian audio-to-audio alignment method. The problem is difficult because, unlike audio-to-score alignment, the underlying music score is not given. To tackle this issue, the underlying musical piece is expressed probabilistically. Then, it is inferred from the input audio signals, assuming that each audio signal plays the same musical piece. Since a music score consists of a relatively few note combinations that get reused throughout the piece, the musical piece is represented as a Markov chain with sparse transition probabilities. Furthermore, the audio signals are represented so that they have different but interrelated tempo trajectories. An experimental evaluation on real-world piano music collection showed a median alignment error of 60 ms, outperforming existing probabilistic methods.

In a real-world use scenario, music alignment is further deteriorated for two more reasons. First, some audio signals may play only a subset of the music score. For example, a user might be interested in aligning between an orchestral piece and a hummed melodic line to the piece. This new kind of alignment is defined as subset music alignment. Second, the room acoustics may vary significantly. For example, an orchestral recording may be recorded in a highly

reverberant hall, while a hummed melodic line to the piece may be recorded in a bedroom. Next two chapters address these two problems.

Chapter 5 presents a subset music alignment method. It is difficult because the input signals play different sequences of note combinations, with some notes that are played in common. To tackle this issue, the proposed method decomposes the input audio into components common to different recordings and those unique to each recording, and simultaneously aligning the audio signals based on the common components. Namely, a hierarchical point process called the hierarchical Dirichlet process (HDP) is used to represent a subset-like relationship in two layers. On the upper layer, one HDP represents each note combination in a piece of music as a subset of all possible notes that are used in the piece. Then, on the lower layer, the other HDP represents the note combinations inside each audio signal as subsets of the note combinations of the top layer HDP. HMMs are used to encode the order of possible notes to play. Evaluations showed that when aligning audio signals that had only one part in common, the proposed method achieved a median alignment of about 200 ms, outperforming existing methods.

Chapter 6 presents a dereverberation method that can be used as a front-end to any alignment method. It is designed to attenuate long-term reverberation (called the late reverberation), which causes past musical notes to smear into the current time instance. Dereverberation is formulated as a deconvolution problem of non-negative autoregressive (AR) process in the power spectrogram domain. Since the order of the AR process is unknown, the Dirichlet process is used to encode a varying number of coefficients that are used. The model is non-conjugate, making inference difficult. To allow inference, a novel inference algorithm based on minorization-maximization is derived. An experimental evaluation showed that even though audio-to-score alignment accuracy degrades under a reverberant environment, the proposed method is capable of recovering the audio-to-score alignment accuracy comparable to that under no reverberation.

Chapter 7 discusses the thesis and presents directions for future research.

注)論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し、合わせて、３,０００字を標準とすること。
論文内容の要旨を英語で記入する場合は、４００～１,１００wordsで作成し
審査結果の要旨は日本語５００～２,０００字程度で作成すること。

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、音楽音響信号を対象としたアライメント、すなわち演奏の音響信号を楽譜に対応づけたり、同一の楽曲を演奏した複数の音響信号を互いに時系列上に対応づける問題に対して、ベイズ推論の枠組みで定式化・実現した研究をまとめたもので、主な成果は以下の通りである。

1. 音響信号と楽譜のアライメント問題をベイズ推論に基づいて定式化した。これにより、楽器の調波構造や楽譜から与えられる事前知識を確率的にモデル化した上で、音色・音量や時間同期に関する不確実性を統一的に扱うことができる。合奏信号を用いた評価実験の結果、従来手法より高い性能（時間誤りの中央値が約30ms）を実現した。

2. 同一の楽曲を演奏した音響信号どうしのアライメント問題に対してもベイズ推論の枠組みを適用した。ここでは楽譜が与えられないので、音響信号セグメントのクラスタで定義される状態のエルゴディックマルコフモデルを用いた。評価実験の結果、ＤＴＷと同等の性能（時間誤りの中央値が60ms）を得た。

3. 上記の枠組みを、各入力信号が楽譜の一部のパートのみを演奏している場合に拡張した。入力信号を、全パートで共通に用いられている要素と各パート固有の要素からなる混合モデルととらえ、これを階層的ディレクレ過程により定式化した。このような問題設定自体に新規性があるが、従来法と比べても有望な結果を得た。

4. 音楽音響信号のアライメントの上で問題となる残響を抑圧する方法をベイズ推論の枠組みで定式化・実現した。これは、原音と残響特性に弱い制約を置いた上で、残響成分を確率的に推論するものである。これにより、長い残響が付加された音楽音響信号に対しても、残響がない場合と同程度の性能でアライメントを行うことが可能になった。

以上のように本論文は、同じ楽曲に対して存在する多数の演奏・音響信号の多様性をベイズ推論という確率的な枠組みで扱う方法を提案したもので、学術上・実用上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、平成27年 2月23日に論文とそれに関連した内容に関する口頭試問を行った結果、合格と認めた。

注)論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　　年　　月　　日以降